Ordered smoothers with exponential weighting

E Chernousova, Yu Golubev, E Krymova

Ordered Smoothers With Exponential Weighting*

Chernousova, E.,† Golubev, Yu.‡ and Krymova, E.§

Abstract

The main goal in this paper is to propose a new approach to deriving oracle inequalities related to the exponential weighting method. The paper focuses on recovering an unknown vector from noisy data with the help of the family of ordered smoothers [12]. The estimators withing this family are aggregated using the exponential weighting method and the aim is to control the risk of the aggregated estimate. Based on natural probabilistic properties of the unbiased risk estimate, we derive new oracle inequalities for mean square risk and show that the exponential weighting permits to improve Kneip's oracle inequality.

1 Introduction and main results

This paper deals with the simple linear model

$$Y_i = \mu_i + \sigma\xi_i, \quad i = 1, 2, \dots, n, \tag{1.1}$$

where ξ is a standard white Gaussian noise, i.e. ξ_i are Gaussian i.i.d. random variables with $\mathbf{E}\xi_i = 0$ and $\mathbf{E}\xi_i^2 = 1$. For the sake of simplicity it is assumed that the noise level $\sigma > 0$ is known.

The goal is to estimate an unknown vector $\mu \in \mathbb{R}^n$ based on the data $Y = (Y_1, \dots, Y_n)^\top$. In this paper, μ is recovered with the help of the following

*This work is partially supported by Laboratory for Structural Methods of Data Analysis in Predictive Modeling, MIPT, RF Government grant, ag. 11.G34.31.0073; and RFBR research projects 13-01-12447 and 13-07-12111.

†Moscow Institute of Physics and Technology, Institutski per. 9, Dolgoprudny, 141700, Russsia, lena-ezhova@rambler.ru

‡Aix Marseille Université, CNRS, LATP, UMR 7353, 13453 Marseille, France and Institute for Information Transmission Problems, Bolshoy Karetny per. 19, Moscow, 127994, Russia, golubev.yuri@gmail.com

§DATADVANCE and Institute for Information Transmission Problems, Bolshoy Karetny per. 19, Moscow, 127994, Russia, ekkrym@gmail.com

family of linear estimates

$$\hat{\mu}_i^h(Y) = h_i Y_i, \ h \in \mathcal{H}, \tag{1.2}$$

where $\mathcal{H}$ is a finite set of vectors in $\mathbb{R}^n$ which will be described later on.

In what follows, the risk of an estimate $\hat{\mu}(Y) = (\hat{\mu}_1(Y), \dots, \hat{\mu}_n(Y))^\top$ is measured by

$$R(\hat{\mu}, \mu) = \mathbf{E}_\mu \|\hat{\mu}(Y) - \mu\|^2,$$

where $\mathbf{E}_\mu$ is the expectation with respect to the measure $\mathbf{P}_\mu$ generated by the observations from (1.1) and $\|\cdot\|$, $\langle \cdot, \cdot \rangle$ stand for the standard norm and inner product in $\mathbb{R}^n$

$$\|x\|^2 = \sum_{i=1}^n x_i^2, \quad \langle x, y \rangle = \sum_{i=1}^n x_i y_i.$$

One can check very easily that the mean square risk of $\hat{\mu}^h(Y)$ is given by

$$R(\hat{\mu}^h, \mu) = \|(1-h) \cdot \mu\|^2 + \sigma^2 \|h\|^2,$$

where $x \cdot y$ denotes the coordinate-wise product of vectors $x, y \in \mathbb{R}^n$, i.e. $z = x \cdot y$, means that $z_i = x_i y_i$, $i = 1, \dots, n$. So, $R(\hat{\mu}^h, \mu)$ depends on $h \in \mathcal{H}$ and one can minimize it choosing properly $h \in \mathcal{H}$. Very often the minimal risk

$$r^{\mathcal{H}}(\mu) = \min_{h \in \mathcal{H}} R(\hat{\mu}^h, \mu)$$

is called the oracle risk.

Obviously, the oracle estimate

$$\mu^*(Y) = h^* \cdot Y, \quad \text{where } h^* = \arg\min_{h \in \mathcal{H}} R(\hat{\mu}^h, \mu),$$

cannot be used since it depends on the unknown vector μ. However, one could try to construct an estimate $\tilde{\mu}^{\mathcal{H}}(Y)$ based on the family of linear estimates $\hat{\mu}^h(Y)$, $h \in \mathcal{H}$, the risk of which is close to the oracle risk. This means that the risk of $\tilde{\mu}^{\mathcal{H}}(Y)$ might be bounded from above by the so-called oracle inequality

$$R(\tilde{\mu}^{\mathcal{H}}, \mu) \le r^{\mathcal{H}}(\mu) + \tilde{\Delta}^{\mathcal{H}}(\mu),$$

which holds uniformly in $\mu \in \mathbb{R}^n$. Heuristically, this inequality assumes that the remainder term $\tilde{\Delta}^{\mathcal{H}}(\mu)$ is negligible with respect to the oracle risk for all $\mu \in \mathbb{R}^n$. In general, such an estimator doesn't exist, but for certain statistical models one can construct an estimator $\tilde{\mu}^{\mathcal{H}}(Y)$ (see, e.g., Theorem 1.1 below) such that:

- $\tilde{\Delta}^{\mathcal{H}}(\mu) \le \tilde{C} r^{\mathcal{H}}(\mu)$ for all $\mu \in \mathbb{R}^n$, where $\tilde{C} > 1$ is a constant.
- $\tilde{\Delta}^{\mathcal{H}}(\mu) \ll r^{\mathcal{H}}(\mu)$ for all μ such that $r^{\mathcal{H}}(\mu) \gg \sigma^2$.

It is also well-known that one can find an estimator with the above properties provided that $\mathcal{H}$ is not very rich. In particular, as shown in [12], this can be done for the so-called *ordered smoothers*. This is why this paper deals with $\mathcal{H}$ containing solely ordered multipliers defined as follows:

Definition 1.1. *$\mathcal{H}$ is a set of ordered multipliers if the following properties hold*

- contracting property*:* $h_i \in [0,1],\ i = 1, \dots, n$ *for all* $h \in \mathcal{H}$,
- decreasing property: $h_{i+1} \le h_i,\ i = 1, \dots, n$ *for all* $h \in \mathcal{H}$,
- totally ordered property*: if for some integer* k *and some* $h, g \in \mathcal{H}$, $h_k < g_k$, *then* $h_i \le g_i$ *for all* $i = 1, \dots, n$.

The totally ordered property means that vectors in $\mathcal{H}$ may be naturally ordered, since for any $h, g \in \mathcal{H}$ there are only two possibilities $h_i \le g_i$ or $h_i \ge g_i$ for all $i = 1, \dots, n$. Therefore the estimators defined by (1.2), where $\mathcal{H}$ is a set of ordered multipliers, are often called ordered smoothers [12].

Note that ordered smoothers are common in statistics. Below we give two basic examples, where these smoothers appear naturally.

Smoothing splines. They are usually used in recovering smooth regression functions $f(x),\ x \in [0,1]$, given the noisy observations $Z = (Z_1, \dots, Z_n)^\top$

$$Z_i = f(X_i) + \varepsilon \xi_i, \quad i = 1, \dots, n, \tag{1.3}$$

where the design points X_i belong to $(0,1)$ and ξ is a standard white Gaussian noise. It is well known that smoothing splines are defined by

$$\hat{f}_\alpha(x, Z) = \arg\min_f \left\{ \frac{1}{n} \sum_{i=1}^{n} [Z_i - f(X_i)]^2 + \alpha \int_0^1 [f^{(m)}(x)]^2 \right\}, \tag{1.4}$$

where $f^{(m)}(\cdot)$ denotes the derivative of order m and $\alpha > 0$ is a smoothing parameter which is usually chosen with the help of the Generalized Cross Validation (see, e.g., [23]).

In order to show that the regression estimation with the help of the smoothing splines is equivalent to the sequence space model (1.1), consider

the Demmler-Reinsch [6] basis $\psi_k(x),\ x \in [0,1],\ k = 1,\dots,n$ having double orthogonality

$$\langle \psi_k, \psi_l \rangle_n = \delta_{kl}, \quad \int_0^1 \psi_k^{(m)}(x)\psi_l^{(m)}(x)\,dx = \delta_{kl}\lambda_k,\ k,l = 1,\dots,n, \tag{1.5}$$

where $\delta_{kl} = 1$ if $k = l$, and $\delta_{kl} = 0$ otherwise. Here and below $\langle u, v \rangle_n$ stands for the inner product

$$\langle u, v \rangle_n = \frac{1}{n}\sum_{i=1}^n u(X_i)v(X_i).$$

Let us assume for definiteness that the eigenvalues λ_k are sorted in ascending order $\lambda_1 \le \ldots \le \lambda_n$.

With this basis, representing the underlying regression function as follows:

$$f(x) = \sum_{k=1}^n \psi_k(x)\mu_k, \tag{1.6}$$

we get from (1.3) and (1.5)

$$Y'_k = \frac{1}{n}\sum_{i=1}^n Z_i\psi_k(X_i) = \mu_k + \frac{\varepsilon}{\sqrt{n}}\xi'_k, \tag{1.7}$$

where $\mu_k = \langle f, \psi_k \rangle_n$ and ξ' is a standard white Gaussian noise. So, substituting (1.6) in (1.4) and using (1.5), we arrive at

$$\hat{f}_\alpha(x, Z) = \sum_{k=1}^n \hat{\mu}_k\psi_k(x),$$

where

$$\hat{\mu} = \arg\min_\mu \left\{ \sum_{k=1}^n [Y'_k - \mu_k]^2 + \alpha\sum_{k=1}^n \lambda_k\mu_k^2 \right\}.$$

It can be seen easily that

$$\hat{\mu}_k = \frac{Y_k}{1 + \alpha\lambda_k}$$

and thus the model (1.3)–(1.4) is equivalent to (1.1)–(1.2) with $\sigma = \varepsilon/\sqrt{n}$ and

$$\mathcal{H} = \left\{ h :\ h_k = \frac{1}{1 + \alpha\lambda_k},\ \alpha \in \mathbb{R}^+ \right\}. \tag{1.8}$$

If we are interested in minimax regression estimates on Sobolev's classes, they can be easily constructed using the statistics from (1.7) and the following set of ordered multipliers

$$\mathcal{H} = \big\{h: \ h_k = \max(1 - \alpha\lambda_k, 0), \ \alpha \in \mathbb{R}^+\big\}.$$

See [17] and [20] for details.

Note that the Demmler-Reinsch basis is a very useful tool for statistical analysis of spline methods. However, in constructing statistical estimates, this basis is rarely used since there are very fast algorithms for computing smoothing splines (see, e.g., [10] and [23]).

Spectral regularizations of large linear models. Very often in linear models, we are interested in estimating $X\theta \in \mathbb{R}^n$ based on the observations

$$Z = X\theta + \sigma\xi, \tag{1.9}$$

where X is a known $n \times p$-matrix, $\theta \in \mathbb{R}^p$ is an unknown vector, and ξ is a standard white Gaussian noise. It is well known that if $X^\top X$ has a large condition number or p is large, then the standard maximum likelihood estimate $X\hat{\theta}^0(Z)$, where

$$\hat{\theta}^0(Z) = \arg\min_\theta \|Z - X\theta\|^2 = (X^\top X)^{-1}X^\top Z$$

may result in a large risk. In particular, if $X^\top X > 0$, then

$$\mathbf{E}\|X\theta - X\hat{\theta}^0\|^2 = \sigma^2 p.$$

When p is large, this risk may be improved with the help of a regularization term. For instance, one can use the Phillps-Tikhonov regularization [22] (often called ridge regression in statistics)

$$\hat{\theta}^\alpha(Z) = \arg\min_\theta\Big\{\|Z - X\theta\|^2 + \alpha\|\theta\|^2\Big\},$$

where $\alpha > 0$ is a smoothing parameter. It can be seen easily that

$$\hat{\theta}^\alpha(Z) = [I + \alpha(X^\top X)^{-1}]^{-1}\hat{\theta}^0(Z).$$

This formula is a particular case of the so-called spectral regularizations defined as follows (see, e.g., [7]):

$$\hat{\theta}^\alpha(Z) = H^\alpha(X^\top X)\hat{\theta}^0(Z),$$

where $H^\alpha(X^\top X)$ is a matrix depending on a smoothing parameter $\alpha \in \mathbb{R}^+$ and $X^\top X$ and admitting the following representation

$$H^\alpha(X^\top X) = \sum_{k=1}^{p} h^\alpha(\lambda_k) e_k e_k^\top$$

where $e_k,\ k = 1, \ldots, p$ and $\lambda_1 \le \lambda_2 \le \ldots \le \lambda_p$ are eigenvectors and eigenvalues of $X^\top X$, and $h^\alpha(\cdot)$ is a function $\mathbb{R}^+ \to [0, 1]$.

The standard way to construct an equivalent model for the spectral regularization method is to make use of the SVD. Note that

$$e_k^* = \frac{X e_k}{\sqrt{\lambda_k}},\ k = 1, \ldots, p$$

is an orthonormal system in $\mathbb{R}^n$. Therefore the observations Z (see (1.9)) admit the following equivalent representation

$$Y_k = \langle e_k^*, Z \rangle = \langle e_k^*, X\theta \rangle + \sigma \xi_k',\ k = 1, \ldots, p, \tag{1.10}$$

where ξ' is a standard white Gaussian noise. Noticing that

$$X\hat{\theta}^\alpha(Z) = X H^\alpha(X^\top X)(X^\top X)^{-1} X^\top Z,$$

we have

$$\begin{aligned}\langle X\hat{\theta}^\alpha(Z), e_k^* \rangle &= \sum_{s=1}^{p} Y_s \langle X H^\alpha(X^\top X)(X^\top X)^{-1} X^\top e_s^*, e_k^* \rangle \\ &= \sum_{s=1}^{p} Y_s \lambda_k \langle H^\alpha(X^\top X)(X^\top X)^{-1} e_s, e_k \rangle = h^\alpha(\lambda_k) Y_k.\end{aligned}$$

Therefore from this equation and (1.10) and we see that the spectral regularization method is equivalent to the statistical model defined by (1.1) and (1.2) with $\mathcal{H} = \{h :\ h_k = h^\alpha(\lambda_k),\ \alpha \in \mathbb{R}^+\}$.

Note that for the Phillps-Tikhonov method we have

$$h^\alpha(\lambda) = \frac{1}{1 + \alpha/\lambda}$$

and it is clear that the corresponding $\mathcal{H}$ is a set of ordered multipliers. Along with this regularization method, the spectral cut-off method and Landweber's iterations (see, e.g., [7] for details) are typical examples of ordered smoothers.

Nowadays, there are a lot of approaches aimed to construct estimates mimicking the oracle risk. At the best of our knowledge, the principal idea in obtaining such estimates goes back to [3] and [15] and related to the method of the unbiased risk estimation [21]. The literature on this approach is so vast that it would be impractical to cite it here. We mention solely the following result by Kneip [12] since it plays an important role in our presentation. Denote by

$$\bar{r}(Y, \hat{\mu}^h) \stackrel{\text{def}}{=} \|Y - \hat{\mu}^h(Y)\|^2 + 2\sigma^2 \sum_{i=1}^{n} h_i - \sigma^2 n, \tag{1.11}$$

the unbiased risk estimate of $\hat{\mu}^h(Y)$.

Theorem 1.1. *Let*

$$\hat{h} = \arg\min_{h \in \mathcal{H}} \bar{r}(Y, \hat{\mu}^h)$$

be a minimizer of the unbiased risk estimate. Then uniformly in $\mu \in \mathbb{R}^n$,

$$\mathbf{E}_\mu \|\hat{h} \cdot Y - \mu\|^2 \le r^{\mathcal{H}}(\mu) + K\sigma^2 \sqrt{1 + \frac{r^{\mathcal{H}}(\mu)}{\sigma^2}}, \tag{1.12}$$

where K is a generic constant.

Another well-known idea to construct a good estimator based on the family $\hat{\mu}^h$, $h \in \mathcal{H}$ is to aggregate the estimates within this family using a held-out sample. Apparently, this approach was firstly developed by Nemirovsky in [16] (see also [11]) and independently by Catoni (see [4] for a summary). Later, the method was extended to several statistical models (see, e.g., [24], [18], [13], [19]).

To overcome the well-know drawbacks of sample splitting, one would like to aggregate estimators using the same observations for constructing estimators and performing the aggregation. This can be done, for instance, with the help of the exponential weighting. The motivation of this method is related to the problem of functional aggregation, see [19]. It has been shown that this method yields rather good oracle inequalities for certain statistical models [14], [5], [19], [1], [2].

In the considered statistical model, the exponential weighting estimate is defined as follows:

$$\bar{\mu}(Y) = \sum_{h \in \mathcal{H}} w_h(Y) \hat{\mu}^h(Y), \tag{1.13}$$

where

$$w_h(Y) = \pi_h \exp\left[-\frac{\bar{r}(Y,\hat{\mu}^h)}{2\beta\sigma^2}\right] \Big/ \sum_{g\in\mathcal{H}} \pi_g \exp\left[-\frac{\bar{r}(Y,\hat{\mu}^g)}{2\beta\sigma^2}\right], \ \beta > 0,$$

$\bar{r}(Y,\hat{\mu}^h)$ is the unbiased risk estimate of $\hat{\mu}^h(Y)$ defined by (1.11), and a priori weights $\{\pi_h,\ h \in \mathcal{H}\}$ are non-negative and such that $\sum_{h\in\mathcal{H}} \pi_h > 0$. Recall that for simplicity, it is assumed here and in what follows that $\mathcal{H}$ is discrete and finite.

It has been shown in [5] that for this method the following oracle inequalities hold.

Theorem 1.2. *Assume that* $\sum_{h\in\mathcal{H}} \pi_h = 1$. *If* $\beta \geq 4$, *then uniformly in* $\mu \in \mathbb{R}^n$

$$\begin{aligned} R(\bar{\mu},\mu) &\leq \min_{\lambda_h\geq 0:\|\lambda\|_1=1} \left\{ \sum_{h\in\mathcal{H}} \lambda_h R(\hat{\mu}^h,\mu) + 2\sigma^2\beta\mathcal{K}(\lambda,\pi) \right\}, \\ R(\bar{\mu},\mu) &\leq \min_{h\in\mathcal{H}} \left\{ R(\hat{\mu}^h,\mu) + 2\sigma^2\beta \log\frac{1}{\pi_h} \right\}, \end{aligned} \tag{1.14}$$

where $\mathcal{K}(\cdot,\cdot)$ *is the Kullback-Leibler divergence and* $\|\cdot\|_1$ *stands for* ℓ_1*-norm, i.e.,*

$$\mathcal{K}(\lambda,\pi) = \sum_{h\in\mathcal{H}} \lambda_h \log\frac{\lambda_h}{\pi_h}, \quad \|\lambda\|_1 = \sum_{h\in\mathcal{H}} |\lambda_h|.$$

Note that for projection methods ($h_k \in \{0,1\}$) this theorem holds for $\beta \geq 2$, see [14].

It is clear that if we want to derive from (1.14) an oracle inequality similar to (1.12), then we have to chose $\pi_h = 1/(\#\mathcal{H})$, where $\#\mathcal{H}$ denotes the cardinality of $\mathcal{H}$, and thus we arrive at

$$R(\bar{\mu},\mu) \leq r^{\mathcal{H}}(\mu) + 2\sigma^2\beta\log(\#\mathcal{H}).$$

This oracle inequality is good only when the cardinality of $\mathcal{H}$ is not very large. If we deal with $\mathcal{H}$ having a very large cardinality like those related to smoothing splines, this inequality is not good. To some extent, this situation may be improved, see Proposition 2 in [5]. However, looking at the oracle inequality in this proposition, one cannot say, unfortunately, that it is always better than (1.12).

The main goal in this paper is to show that for the exponential weighting method one can obtain an oracle inequality with a smaller remainder term than the one in Theorem 1.1, Equation (1.12).

In order to attain this goal and to cover $\mathcal{H}$ with both small and large cardinality, we make use of the special prior weights defined as follows:

$$\pi_h \stackrel{\text{def}}{=} 1 - \exp\left\{-\frac{\|h^+\|_1 - \|h\|_1}{\beta}\right\}. \tag{1.15}$$

Here

$$h^+ = \min\{g \in \mathcal{H} : g > h\}$$

and $\pi_{h^{\max}} = 1$, where $h^{\max}$ is the maximal multiplier in $\mathcal{H}$.

Along with these weights we will need also the following condition:

Condition 1.1. *There exists a constant $K_\circ \in (0, \infty)$ such that*

$$\|h\|^2 - \|g\|^2 \geq K_\circ\big(\|h\|_1 - \|g\|_1\big) \tag{1.16}$$

for all $h \geq g$ from $\mathcal{H}$.

The next theorem, yielding an upper bound for the mean square risk of $\bar{\mu}(Y)$ defined by (1.13), is the main result of this paper.

Theorem 1.3. *Assume that $\mathcal{H}$ is a set of ordered multipliers, $\beta \geq 4$, and Condition 1.1 holds. Then, uniformly in $\mu \in \mathbb{R}^n$,*

$$\mathbf{E}_\mu\|\bar{\mu}(Y) - \mu\|^2 \leq r^{\mathcal{H}}(\mu) + 2\beta\sigma^2 \log\left[C\left(1 + \frac{r^{\mathcal{H}}(\mu)}{\sigma^2}\right)\right]. \tag{1.17}$$

Here and in what follows $C = C(K_\circ, \beta)$ denotes strictly positive and bounded constants depending on $K_\circ, \beta$.

We finish this section with some remarks regarding this theorem.

Remark 1. The condition $\beta \geq 4$ may be improved when the ordered multipliers $h \in \mathcal{H}$ take only two values 0 and 1. In this case it is sufficient to assume that $\beta \geq 2$ (see [9]).

Remark 2. Usually Condition 1.1 may be checked rather easily. For instance, for smoothing splines and equidistant design, the set of ordered multipliers is given by (1.8) and this condition follows from the well-known asymptotic formula $\lambda_k \asymp (\pi k)^{2m}$ for large k (see [6] for details). Heuristically, for small α and large n we have

$$\begin{aligned}\|h^\alpha\|^2 = \sum_{k=1}^n \frac{1}{(1+\alpha\lambda_k)^2} &\approx \sum_{k=1}^n \frac{1}{[1+\alpha(\pi k)^{2m}]^2} \\ &\approx \frac{1}{\pi\alpha^{1/(2m)}}\int_0^\infty \frac{1}{[1+x^{2m}]^2}\,dx\end{aligned}$$

and

$$\|h^{\alpha}\|_1 = \sum_{k=1}^{n} \frac{1}{1+\alpha\lambda_k} \approx \sum_{k=1}^{n} \frac{1}{1+\alpha(\pi k)^{2m}} \approx \frac{1}{\pi\alpha^{1/(2m)}} \int_0^{\infty} \frac{1}{1+x^{2m}}\, dx. \tag{1.18}$$

With these equations Condition 1.1 becomes obvious. A rigorous proof of (1.16) is based on a non-asymptotic version of these arguments. It is technical but unfortunately cumbersome and therefore, in order not to overload the paper, we omit it.

For spectral regularizations Condition 1.1 is obvious for the spectral cut-off method. For the Phillps-Tikhonov method the proof is more involved but similar to the one for splines. Note, however, that in this case the following condition

$$\lambda_s \geq K s^q, \text{ for some } q > 1$$

is required.

Remark 3. In practice, the multipliers in $\mathcal{H}$ are often chosen so that

$$\|h^{+}\|_1 = (1+\epsilon)\|h\|_1$$

with some small $\epsilon > 0$ and the initial condition $\|h^{\min}\|_1 = 1$, where $h^{\min}$ is the minimal multiplier in $\mathcal{H}$. In this case one may chose the a priori weights $\pi_h = 1$ since π_h from (1.15) are strictly bounded from below and it can be checked very easily that Lemma 2.1 holds true for $\pi_h = 1$. This means in particular that if the smoothing parameter α in smoothing splines (1.4) takes values $\{(1+\epsilon)^{-2mk}, k = 0, 1, \ldots\}$, then $\pi_h = 1$ may be used in the exponential weighting (see (1.18)).

Remark 4. In contrast to Proposition 2 in [5], the remainder term in (1.17) does depend neither on the cardinality of $\mathcal{H}$ nor n. It has the same structure as Kneip's oracle inequality in Theorem 1.1.

Remark 5. Comparing (1.17) with (1.12), we see that when

$$\frac{r^{\mathcal{H}}(\mu)}{\sigma^2} \approx 1,$$

then the remainder terms in (1.12) and (1.17) have the same order, namely, σ^2. However, when

$$\frac{r^{\mathcal{H}}(\mu)}{\sigma^2} \gg 1,$$

we get

$$2\beta\sigma^2 \log\left[C\left(1+\frac{r^{\mathcal{H}}(\mu)}{\sigma^2}\right)\right] \ll K\sigma^2\sqrt{1+\frac{r^{\mathcal{H}}(\mu)}{\sigma^2}},$$

thus showing that the upper bound for the remainder term in the oracle inequality related to the exponential weighting is better than the one in Theorem 1.1.

Remark 6. To compare the actual remainder terms in (1.17) and (1.12) and to find out what β is good from a practical viewpoint, a numerical experiment has been carried out. The goal in this experiment is to compare the exponential weighting methods for $\beta = \{0, 1, 2, 4\}$ combined with the cubic smoothing splines for the equidistant design. In order to simulate these splines, the following family of ordered multipliers was used

$$\mathcal{H} = \left\{h : h_k = \frac{1}{1+[\alpha(k-1)]^4},\ \alpha > 0\right\}.$$

The motivation of this family is due to the well-known asymptotic formula $\lambda_k \asymp (\pi k)^4,\ k \to \infty$.

The simulations are organized as follows. For given $A \in [0, 300]$, 100000 replications of the observations

$$Y_k = \mu_k(A) + \xi_k,\ k = 1, \dots, 400$$

are generated. Here $\mu(A) \in \mathbb{R}^{400}$ is a Gaussian vector with independent components and

$$\mathbf{E}\mu_k(A) = 0, \quad \mathbf{E}\mu_k^2(A) = A\exp\left(-\frac{k^2}{2\Sigma^2}\right).$$

Next, the mean oracle risk

$$\bar{r}^{\mathcal{H}}(A) = \mathbf{E}\min_{h\in\mathcal{H}}\left\{\|(1-h)\cdot\mu(A)\|^2 + \|h\|^2\right\}$$

and the mean excess risk

$$\bar{\Delta}_\beta(A) = \mathbf{E}\|\mu(A) - \bar{\mu}(Y)\|^2 - \bar{r}^{\mathcal{H}}(A),$$

were computed with the help of the Monte-Carlo method. Finally, the data $\{\bar{r}^{\mathcal{H}}(A), \bar{\Delta}_\beta(A), A \in [0, 300]\}$ are plotted on Figure 1 to illustrate graphically the remainder term $\Delta_\beta(r^{\mathcal{H}}) = \mathbf{E}_\mu\|\bar{\mu} - \mu\|^2 - r^{\mathcal{H}}(\mu)$.

Looking at this picture we see that there is no universal β minimizing the excess risk uniformly in μ. However, it seems to us that a reasonable

choice is $\beta \approx 1$, but unfortunately, good oracle inequalities are not available for this case. Note also that the exponential weighting can provide only a moderate improvement of the risk compared to the classical unbiased risk estimation ($\beta = 0$). All methods demonstrate almost similar statistical performance. However, when $r^{\mathcal{H}}(\mu)/\sigma^2$ is not large, the exponential weighting works usually better.

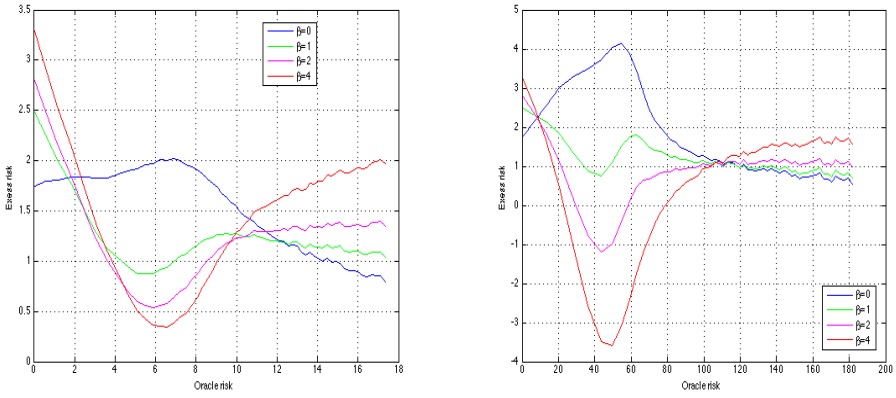

Figure 1: Mean excess risks $\bar{\Delta}_\beta(\bar{r}^{\mathcal{H}})$ (left panel $\Sigma = 5$; right panel $\Sigma = 50$).

2 Proofs

The proof of Theorem 1.3 is based on a combination of methods for deriving oracle inequalities proposed in [14] and [9]. We indicate here solely the main steps in the proof, all details are given below. With the help of Stein's formula for the unbiased risk estimate it can be shown similar to [14] that for $\beta \geq 4$

$$\begin{aligned}
\mathbf{E}_\mu \|\bar{\mu} - \mu\|^2 \leq & \mathbf{E}_\mu \sum_{h \in \mathcal{H}} w_h(Y) \bar{r}(Y, \hat{\mu}^h) \\
\leq & r^{\mathcal{H}}(\mu) + 2\beta\sigma^2 \mathbf{E}_\mu \sum_{h \in \mathcal{H}} w_h(Y) \log \frac{\pi_h}{w_h(Y)} \\
& - 2\beta\sigma^2 \mathbf{E}_\mu \log \left\{ \sum_{h \in \mathcal{H}} \pi_h \exp \left[- \frac{\bar{r}(Y, \hat{\mu}^h) - \bar{r}(Y, \hat{\mu}^{\hat{h}})}{2\beta\sigma^2} \right] \right\},
\end{aligned} \tag{2.1}$$

where $\hat{h} = \arg\min_{h \in \mathcal{H}} \bar{r}(Y, \hat{\mu}^h)$.

To control the right-hand side at this equation, we make use of the ordering property of estimates $\hat{\mu}^h,\ h \in \mathcal{H}$. First, we check using (**??**) that if π_h is defined by (1.15), then

$$\sum_{h\in\mathcal{H}} \pi_h \exp\left[-\frac{\bar{r}(Y,\hat{\mu}^h) - \bar{r}(Y,\hat{\mu}^{\hat{h}})}{2\beta\sigma^2}\right] \geq \sum_{h\geq\hat{h}} \pi_h \exp\left[-\frac{\bar{r}(Y,\hat{\mu}^h) - \bar{r}(Y,\hat{\mu}^{\hat{h}})}{2\beta\sigma^2}\right] \geq 1,$$

and so, the last term in Equation (2.1) is always negative.

The most difficult and delicate part of the proof is related to the average Kullback-Leibler divergence

$$\mathbf{E}_\mu \sum_{h\in\mathcal{H}} w_h(Y) \log \frac{w_h(Y)}{\pi_h}.$$

To compute a good lower bound for this value, we follow the approach proposed in [9]. The main idea here is to make use of the following property of the unbiased risk estimate: for any sufficiently small $\epsilon < 1$, there exists $\hat{h}^\epsilon$ depending on Y such that with probability 1, for all $h \geq \hat{h}^\epsilon$

$$\bar{r}(Y,\hat{\mu}^h) - \bar{r}(Y,\hat{\mu}^{\hat{h}}) \geq 2\beta\sigma^2\epsilon\big[\|h\|^2 - \|\hat{h}\|^2\big] + 2\beta\sigma^2.$$

This equation means that $w_h(Y)$ are exponentially decreasing for large h. With this property we obtain the following entropy bound (see Lemma 2.3 below)

$$\sum_{h\in\mathcal{H}} w_h(Y) \log \frac{\pi_h}{w_h(Y)} \leq \log\left[\sum_{h\leq\hat{h}^\epsilon} \pi_h + \frac{C}{\epsilon}\exp\left(\frac{C}{\epsilon}\right)\right].$$

The rest of the proof is routine. It follows from (1.16) and (1.15) (see Lemma 2.1 below) that

$$\sum_{h\leq\hat{h}^\epsilon} \pi_h \leq 1 + \frac{\|\hat{h}^\epsilon\|^2}{K_\circ\beta}.$$

More cumbersome probabilistic technique is required to prove the following upper bound (see Lemma 2.5):

$$\sqrt{\mathbf{E}_\mu\|\hat{h}^\epsilon\|^2} \leq \sqrt{\frac{r^{\mathcal{H}}(\mu)}{(1-2\beta\epsilon)\sigma^2}} + \frac{\sqrt{1+2\beta}}{1-2\beta\epsilon}\sqrt{K}.$$

Finally, combining the above equations, we arrive at (1.17).

2.1 Auxiliary facts

The next lemma collects some useful facts about a priori weights defined by (1.15). Let

$$\mathcal{D}^h = \{g \in \mathcal{H} : \|h\|_1 \le \|g\|_1 \le \|h\|_1 + 1\}.$$

Lemma 2.1. *Under Condition 1.1, for any $h \in \mathcal{H}$, the following assertions hold:*

$$\sum_{g \ge h} \pi_g \exp\left\{-\frac{\|g\|_1}{\beta}\right\} = \exp\left\{-\frac{\|h\|_1}{\beta}\right\}, \tag{2.2}$$

$$\sum_{g \le h} \pi_g \le 1 + \frac{\|h\|^2}{K_\circ \beta}, \tag{2.3}$$

$$\sum_{g \in \mathcal{D}^h} \pi_g \le 1 + \frac{1}{\beta}, \tag{2.4}$$

$$\sum_{g \in \mathcal{D}^h} \pi_g \ge \frac{1}{2\beta} \exp\left(-\frac{1}{\beta}\right). \tag{2.5}$$

Proof. Denote for brevity

$$S(h) = \sum_{g \ge h} \pi_g \exp\left\{-\frac{\|g\|_1 - \|h\|_1}{\beta}\right\}.$$

Then we have

$$\begin{aligned} S(h) - S(h^+) &= \pi_h + \exp\left\{-\frac{\|h^+\|_1 - \|h\|_1}{\beta}\right\} \\ &\times \sum_{g \ge h^+} \pi_g \exp\left\{-\frac{\|g\|_1 - \|h^+\|_1}{\beta}\right\} - \sum_{g \ge h^+} \pi_g \exp\left\{-\frac{\|g\|_1 - \|h^+\|_1}{\beta}\right\} \\ &= \pi_h - \left\{1 - \exp\left[-\frac{\|h^+\|_1 - \|h\|_1}{\beta}\right]\right\} S(h^+). \end{aligned}$$

Therefore in view of the definition of π_h, it is clear that if $S(h^{\max}) = 1$, then $S(h) = S(h^+)$, thus proving (2.2).

To prove (2.3), note that

$$\pi_g \le \frac{\|g^+\|_1 - \|g\|_1}{\beta}. \tag{2.6}$$

This inequality follows from (1.15) and from the inequality inequality $1 - \exp(-x) \le x$. Hence, by Condition (1.16) we obtain

$$\begin{aligned}\sum_{g\le h}\pi_g &\le 1+\sum_{g\le h^-}\pi_g \le 1+\frac{1}{\beta}\sum_{g\le h^-}\big[\|g^+\|_1-\|g\|_1\big]\\ &= 1+\frac{\|h\|_1-\|h^{\min}\|_1}{\beta}\le 1+\frac{\|h\|^2-\|h^{\min}\|^2}{K_\circ\beta},\end{aligned}$$

where $h^{\min}$ is the minimal element in $\mathcal{H}$.

The same arguments can be used in proving (2.4).

In order to check (2.5), denote by g_h be the maximal element in $\mathcal{D}^h$. Then there are two possibilities

- $\|g_h\|_1 \le \|h\|_1 + 1/2$,
- $\|g_h\|_1 > \|h\|_1 + 1/2$.

In the first case, $\|g_h^+\|_1 - \|g_h\|_1 \ge 1/2$, and thus by

$$\sum_{g\in\mathcal{D}^h}\pi_g \ge \pi_{g_h} \ge 1-\exp\bigg(-\frac{\|g_h^+\|_1-\|g_h\|_1}{\beta}\bigg) \ge 1-\exp\bigg(-\frac{1}{2\beta}\bigg). \tag{2.7}$$

In the second case, where $\|h\|_1 + 1/2 < \|g_h\|_1 \le \|h\|_1 + 1$, we obtain by a Taylor expansion that for any $g < g_h$

$$\pi_g \ge \frac{\|g^+\|_1-\|g\|_1}{\beta}\exp\bigg(-\frac{\|g_h\|_1-\|h\|_1}{\beta}\bigg) \ge \frac{\|g^+\|_1-\|g\|_1}{\beta}\exp\bigg(-\frac{1}{\beta}\bigg),$$

and thus

$$\sum_{g\in\mathcal{D}^h}\pi_g \ge \frac{\|g_h\|_1-\|h\|_1}{\beta}\exp\bigg(-\frac{1}{\beta}\bigg) \ge \frac{1}{2\beta}\exp\bigg(-\frac{1}{\beta}\bigg).$$

This equation together with (2.7) ensures (2.5) since it can be checked with a simple algebra that

$$\frac{1}{2\beta}\exp\bigg(-\frac{1}{\beta}\bigg) \le 1-\exp\bigg(-\frac{1}{2\beta}\bigg),\ \beta>0. \quad \blacksquare$$

The following lemma is a cornerstone in the proof of Theorem 1.3.

Lemma 2.2. *For $\beta \geq 4$ the risk of $\bar{\mu}(Y)$ is bounded from above as follows:*

$$\mathbf{E}_\mu \|\bar{\mu}(Y) - \mu\|^2 \leq \mathbf{E}_\mu \sum_{h \in \mathcal{H}} w_h(Y) \bar{r}(Y, \hat{\mu}^h).$$

Proof. It is based essentially on the method proposed in [14]. Unfortunately, we cannot use directly Corollary 2 in [14] because it holds only for h_k, $k = 1, \ldots, n$ taking values 0 and 1. In the case of ordered smoothers h_k belongs to the interval $[0, 1]$ and we will see below that this fact results in the condition $\beta \geq 4$.

Recall that the unbiased risk estimates for $\bar{\mu}_i(Y)$ and $\bar{\mu}_i^h(Y)$ are computed as follows (see, e.g. [21])

$$\begin{aligned} \bar{r}(Y_i, \bar{\mu}_i) &= [\bar{\mu}_i(Y) - Y_i]^2 + 2\sigma^2 \frac{\partial \bar{\mu}_i(Y)}{\partial Y_i} - \sigma^2, \\ \bar{r}(Y_i, \hat{\mu}_i^h) &= [\hat{\mu}_i^h(Y) - Y_i]^2 + 2\sigma^2 h_i - \sigma^2. \end{aligned} \tag{2.8}$$

Since $\sum_{h \in \mathcal{H}} w_h = 1$, we have

$$\begin{aligned} &[\bar{\mu}_i(Y) - Y_i]^2 = \sum_{h \in \mathcal{H}} w_h(Y) [\bar{\mu}_i(Y) - Y_i]^2 \\ &= \sum_{h \in \mathcal{H}} w_h(Y) [\bar{\mu}_i(Y) - \hat{\mu}_i^h(Y) + \hat{\mu}_i^h(Y) - Y_i]^2 \\ &= \sum_{h \in \mathcal{H}} w_h(Y) [\bar{\mu}_i(Y) - \hat{\mu}_i^h(Y)]^2 + \sum_{h \in \mathcal{H}} w_h(Y) [\hat{\mu}_i^h(Y) - Y_i]^2 \\ &\quad + 2 \sum_{h \in \mathcal{H}} w_h(Y) [\bar{\mu}_i(Y) - \hat{\mu}_i^h(Y)][\hat{\mu}_i^h(Y) - Y_i] \\ &= \sum_{h \in \mathcal{H}} w_h(Y) [\bar{\mu}_i(Y) - \hat{\mu}_i^h(Y)]^2 + \sum_{h \in \mathcal{H}} w_h(Y) [\hat{\mu}_i^h(Y) - Y_i]^2 \\ &\quad + 2 \sum_{h \in \mathcal{H}} w_h(Y) [\bar{\mu}_i(Y) - \hat{\mu}_i^h(Y)][\hat{\mu}_i^h(Y) - \bar{\mu}_i(Y) + \bar{\mu}_i(Y) - Y_i] \\ &= - \sum_{h \in \mathcal{H}} w_h(Y) [\bar{\mu}_i(Y) - \hat{\mu}_i^h(Y)]^2 + \sum_{h \in \mathcal{H}} w_h(Y) [\hat{\mu}_i^h(Y) - Y_i]^2. \end{aligned} \tag{2.9}$$

From the definition of $\bar{\mu}(Y)$ we obviously get

$$\frac{\partial \bar{\mu}_i(Y)}{\partial Y_i} = \sum_{h \in \mathcal{H}} w_h(Y) \frac{\partial \hat{\mu}_i^h(Y)}{\partial Y_i} + \sum_{h \in \mathcal{H}} \frac{\partial w_h(Y)}{\partial Y_i} \hat{\mu}_i^h(Y)$$

and combining this equation with (2.9) (see also (2.8)), we arrive at

$$\begin{aligned}
\bar{r}(Y_i,\bar{\mu}_i) &= [\bar{\mu}_i(Y)-Y_i]^2 + 2\sigma^2\frac{\partial\bar{\mu}_i(Y)}{\partial Y_i} - \sigma^2 \\
&= \sum_{h\in\mathcal{H}} w_h(Y)\Big\{[\hat{\mu}_i^h(Y)-Y_i]^2 + 2\sigma^2\frac{\partial\hat{\mu}_i^h(Y)}{\partial Y_i} - \sigma^2 \\
&\quad - [\bar{\mu}_i(Y)-\hat{\mu}_i^h(Y)]^2 + 2\sigma^2\frac{\partial\log(w_h(Y))}{\partial Y_i}\hat{\mu}_i^h(Y)\Big\} \\
&= \sum_{h\in\mathcal{H}} w_h(Y)\bar{r}(Y_i,\hat{\mu}_i^h) + \\
&\quad + \sum_{h\in\mathcal{H}} w_h(Y)\Big\{-[\bar{\mu}_i(Y)-\hat{\mu}_i^h(Y)]^2 + 2\sigma^2\frac{\partial\log[w_h(Y)]}{\partial Y_i}\hat{\mu}_i^h(Y)\Big\} \\
&= \sum_{h\in\mathcal{H}} w_h(Y)\bar{r}(Y_i,\hat{\mu}_i^h) + \sum_{h\in\mathcal{H}} w_h(Y)\Big\{-[\bar{\mu}_i(Y)-\hat{\mu}_i^h(Y)]^2 \\
&\quad + 2\sigma^2\frac{\partial\log[w_h(Y)]}{\partial Y_i}[\hat{\mu}_i^h(Y)-\bar{\mu}_i(Y)]\Big\}.
\end{aligned} \tag{2.10}$$

In deriving the above equation it was used that $\sum_{h\in\mathcal{H}} w_h(Y)=1$ and hence

$$\sum_{h\in\mathcal{H}}\frac{\partial w_h(Y)}{\partial Y_i} = \sum_{h\in\mathcal{H}} w_h(Y)\frac{\partial\log w_h(Y)}{\partial Y_i} = 0.$$

To control the second sum at the right hand side of (2.10) we make use of the following equation

$$\log[w_h(Y)] = -\frac{\bar{r}(Y,\hat{\mu}^h)}{2\beta\sigma^2} + \log(\pi_h) - \log\Big\{\sum_{g\in\mathcal{H}}\pi_g\exp\Big[-\frac{\bar{r}(Y,\hat{\mu}^g)}{2\beta\sigma^2}\Big]\Big\}. \tag{2.11}$$

Therefore

$$\begin{aligned}
&\sum_{h\in\mathcal{H}} w_h(Y)\frac{\partial\log w_h(Y)}{\partial Y_i}[\hat{\mu}_i^h(Y)-\bar{\mu}_i(Y)] \\
&\qquad = -\frac{1}{2\beta\sigma^2}\sum_{h\in\mathcal{H}} w_h(Y)\frac{\partial\bar{r}(Y,\hat{\mu}^h)}{\partial Y_i}[\hat{\mu}_i^h(Y)-\bar{\mu}_i(Y)].
\end{aligned}$$

Substituting in the above equation (see (1.11))

$$\frac{\partial\bar{r}(Y_i,\hat{\mu}_i^h)}{\partial Y_i} = 2(1-h_i)^2 Y_i,$$

we obtain

$$\begin{aligned}\sum_{h\in\mathcal{H}} w_h(Y)\frac{\partial \log w_h(Y)}{\partial Y_i}[\hat{\mu}_i^h(Y)-\bar{\mu}_i(Y)]\\ =-\frac{1}{\beta\sigma^2}Y_i^2\sum_{h\in\mathcal{H}} w_h(Y)[h_i-1]^2[h_i-\bar{h}_i],\end{aligned} \tag{2.12}$$

where

$$\bar{h}_i=\sum_{h\in\mathcal{H}} w_h(Y)h_i.$$

Next noticing that

$$(1-h_i)^2=(1-\bar{h}_i)^2+(\bar{h}_i-h_i)^2+2(1-\bar{h}_i)(\bar{h}_i-h_i),$$

we have

$$\begin{aligned}-Y_i^2\sum_{h\in\mathcal{H}} w_h(Y)(h_i-1)^2(h_i-\bar{h}_i) &= Y_i^2(1-\bar{h}_i)^2\sum_{h\in\mathcal{H}} w_h(Y)(\bar{h}_i-h_i)\\ &\quad +Y_i^2\sum_{h\in\mathcal{H}} w_h(Y)(\bar{h}_i-h_i)^2(\bar{h}_i-h_i+2-2\bar{h}_i)\\ &= 2Y_i^2\sum_{h\in\mathcal{H}} w_h(Y)(\bar{h}_i-h_i)^2\left(1-\frac{h_i+\bar{h}_i}{2}\right)\\ &\le 2\sum_{h\in\mathcal{H}} w_h(Y)[\bar{\mu}_i(Y)-\hat{\mu}_i^h(Y)]^2.\end{aligned}$$

Combining this equation with (2.9)–(2.12), we finish the proof. ■

Lemma 2.3. *Suppose* $\{q_h\le 1,\ h\in\mathcal{H}\}$ *is a nonnegative sequence such that for all* $h\ge\tilde{h}$

$$q_h\le\exp\Big\{-\gamma\big[\|h\|_1-\|\tilde{h}\|_1\big]-1\Big\},\quad \gamma>0.$$

Let

$$W_h=\pi_h q_h\left[\sum_{g\in\mathcal{H}}\pi_g q_g\right]^{-1}$$

and $\mathcal{G}$ *be a subset in* $\mathcal{H}$. *Then*

$$H(W,\pi)\stackrel{\text{def}}{=}\sum_{h\in\mathcal{H}} W_h\log\frac{\pi_h}{W_h}\le\log\left[\sum_{h\le\tilde{h}}\pi_h+\exp[R(\gamma)]\right],$$

where

$$R(\gamma) = \log\biggl[\frac{2}{\gamma\beta\mathrm{e}} + \sum_{h\in\mathcal{G}}\pi_h\biggr] + \biggl[\sum_{h\in\mathcal{G}}\pi_h q_h\biggr]^{-1}\biggl[\frac{8}{\gamma\beta\mathrm{e}} + \sum_{h\in\mathcal{G}}\pi_h\biggr]. \tag{2.13}$$

Proof. Decompose $\mathcal{H}$ onto two subsets

$$\mathcal{Q} = \{h \geq \tilde{h}\} \cup \mathcal{G}, \quad \mathcal{P} = \mathcal{H} \setminus \mathcal{Q}$$

and denote for brevity

$$P = \sum_{h\in\mathcal{P}}\pi_h q_h, \quad Q = \sum_{h\in\mathcal{Q}}\pi_h q_h.$$

By convexity of $\log(x)$ we obtain

$$\begin{aligned} H(W,\pi) &= \frac{P}{P+Q}\sum_{h\in\mathcal{P}}\frac{\pi_h q_h}{P}\log\frac{(P+Q)/P}{q_h/P} \\ &\quad + \frac{Q}{P+Q}\sum_{h\in\mathcal{Q}}\frac{\pi_h q_h}{Q}\log\frac{(P+Q)/Q}{q_h/Q} \\ &\leq \frac{P}{P+Q}\log\frac{P+Q}{P} + \frac{Q}{P+Q}\log\frac{P+Q}{Q} + \frac{P}{P+Q}\log\biggl(\sum_{h\in\mathcal{P}}\pi_h\biggr) \\ &\quad + \frac{1}{P+Q}\biggl[\sum_{h\in\mathcal{Q}}\pi_h q_h\log\frac{1}{q_h} + Q\log(Q)\biggr]. \end{aligned} \tag{2.14}$$

Next, note that $f(x) = x\log(1/x)$ is an increasing function, when $x \in (0, \mathrm{e}^{-1}]$, and $\max_{x\in[0,1]} f(x) = \mathrm{e}^{-1}$. Therefore, if

$$q_h \leq \exp\bigl[-\gamma(\|h\|_1 - \|\tilde{h}\|_1) - 1\bigr],$$

then

$$q_h\log\frac{1}{q_h} \leq \exp\bigl[-\gamma(\|h\|_1 - \|\tilde{h}\|_1) - 1\bigr]\bigl[\gamma(\|h\|_1 - \|\tilde{h}\|_1) + 1\bigr],$$

and we get

$$\begin{aligned} \sum_{h\in\mathcal{Q}}\pi_h q_h\log\frac{1}{q_h} \leq \frac{1}{\mathrm{e}}\sum_{h\in\mathcal{G}}\pi_h + \frac{1}{\mathrm{e}}\sum_{h\geq\tilde{h}}\pi_h\exp\bigl[-\gamma(\|h\|_1 - \|\tilde{h}\|_1)\bigr] \\ \times\bigl[\gamma(\|h\|_1 - \|\tilde{h}\|_1) + 1\bigr]. \end{aligned}$$

We continue this equation with (2.6) as follows:

$$\sum_{h\in\mathcal{Q}} \pi_h q_h \log\frac{1}{q_h} \le \frac{1}{\mathrm{e}}\sum_{h\in\mathcal{G}}\pi_h + \frac{1}{\beta\mathrm{e}}\sum_{h\ge\tilde h}\exp\bigl[-\gamma(\|h\|_1-\|\tilde h\|_1)\bigr] \times\bigl[\gamma\bigl(\|h\|_1-\|\tilde h\|_1\bigr)+1\bigr]\bigl(\|h^+\|_1-\|h\|_1\bigr). \tag{2.15}$$

In order to bound from above the right-hand side at this equation, let us index the elements in $\{h\in\mathcal{H}:\ h\ge\tilde h\}$ denoting them as $\{h_k,\ k\ge 0\}$, so that $\|h_{k+1}\|_1\ge\|h_k\|_1$ and $h_0=\tilde h$. Then, denoting for brevity

$$S_i = \|h_i\|_1 - \|\tilde h\|_1,$$

we can rewrite the sum at the right hand side in (2.15) as follows:

$$\sum_{h\ge\tilde h}\exp\bigl[-\gamma(\|h\|_1-\|\tilde h\|_1)\bigr]\bigl[\gamma\bigl(\|h\|_1-\|\tilde h\|_1\bigr)+1\bigr]\bigl(\|h^+\|_1-\|h\|_1\bigr) = \sum_{i\ge0}\exp\bigl[-\gamma S_i\bigr]\bigl[\gamma S_i+1\bigr](S_{i+1}-S_i). \tag{2.16}$$

To bound from above the right hand side at this equation, let us check that

$$\max_{S_k,k\ge1}\sum_{i\ge0}\exp\bigl[-\gamma S_i\bigr](S_{i+1}-S_i)\le\frac{2}{\gamma}, \tag{2.17}$$

where max is computed over all nondecreasing sequences. Solving the equation

$$\frac{\partial}{\partial S_k}\sum_{i\ge0}\exp\bigl[-\gamma S_i\bigr][S_{i+1}-S_i]_+ = 0,$$

we obtain with a simple algebra

$$S_{k+1}-S_k = \frac{\exp\bigl[\gamma(S_k-S_{k-1})\bigr]-1}{\gamma}.$$

Hence

$$\exp(-\gamma S_k)(S_{k+1}-S_k) = \frac{\exp[-\gamma S_{k-1}]-\exp[-\gamma S_k]}{\gamma}$$

and summing up these equations, we arrive at (2.17). Next notice that for any $z>0$ we have $(1+x)\le z\exp(x/z)$. With this inequality and (2.17) we obtain for any $z>1$

$$\sum_{i\ge0}\exp\bigl[-\gamma S_i\bigr]\bigl[\gamma S_i+1\bigr](S_{i+1}-S_i)\le z\sum_{i\ge0}\exp[-\gamma(1-1/z)S_i](S_{i+1}-S_i) \le\frac{2z^2}{(z-1)\gamma}.$$

It can be seen easily that the minimun of the right hand side at this equation is attained at $z = 2$. So, we obtain

$$\sum_{i\geq 0} \exp\big[-\gamma S_i\big]\big[\gamma S_i + 1\big](S_{i+1} - S_i) \leq \frac{8}{\gamma}. \tag{2.18}$$

With Equations (2.15)–(2.18) we get

$$\sum_{h\in\mathcal{Q}} \pi_h q_h \log\frac{1}{q_h} \leq \frac{8}{\gamma\beta\mathrm{e}} + \frac{1}{\mathrm{e}}\sum_{h\in\mathcal{G}} \pi_h \tag{2.19}$$

and similarly

$$Q = \sum_{h\in\mathcal{Q}} \pi_h q_h \leq \frac{2}{\gamma\beta\mathrm{e}} + \sum_{h\in\mathcal{G}} \pi_h. \tag{2.20}$$

Therefore

$$\log(Q) \leq \log\left[\frac{2}{\gamma\beta\mathrm{e}} + \sum_{h\in\mathcal{G}} \pi_h\right]. \tag{2.21}$$

Next, denoting for brevity

$$x = \frac{Q}{P+Q},$$

and using (2.19) and (2.21), we arrive at

$$\begin{aligned} H(W,\pi) \leq \max_{x\in[0,1]}\Big\{ & -x\log(x) - (1-x)\log(1-x) \\ & + (1-x)\log\left(\sum_{h\in\mathcal{P}} \pi_h\right) + x\rho\Big\}, \end{aligned} \tag{2.22}$$

where

$$\begin{aligned} \rho &\stackrel{\text{def}}{=} \log(Q) + \frac{1}{Q}\sum_{h\in\mathcal{Q}} \pi_h q_h \log\frac{1}{q_h} \\ &\leq \log\left[\frac{2}{\gamma\beta\mathrm{e}} + \sum_{h\in\mathcal{G}} \pi_h\right] + \left[\sum_{h\in\mathcal{G}} \pi_h q_h\right]^{-1}\left[\frac{8}{\gamma\beta\mathrm{e}} + \sum_{h\in\mathcal{G}} \pi_h\right] = R(\gamma). \end{aligned}$$

It is seen easily that the minimizer x^* of the right-hand side at (2.22) is a solution to the following equation

$$\log\frac{1-x^*}{x^*} = \log\left(\sum_{h\in\mathcal{P}} \pi_h\right) - \rho$$

and thus

$$x^* = \left\{1 + \left(\sum_{h\in\mathcal{P}} \pi_h\right) \exp(-\rho)\right\}^{-1}.$$

Therefore from (2.22) we get

$$\begin{aligned} H(W,\pi) \le & \log\left(\sum_{h\in\mathcal{P}} \pi_h\right) - \log(1-x^*) \\ & - x^*\left[\log\frac{x^*}{1-x^*} + \log\left(\sum_{h\in\mathcal{P}} \pi_h\right) - \rho\right] \\ & = \log\left(\sum_{h\in\mathcal{P}} \pi_h\right) - \log(1-x^*) = \log\left[\sum_{h\in\mathcal{P}} \pi_h + \exp(\rho)\right] \\ & \le \log\left[\sum_{h<\tilde{h}} \pi_h + \exp(\rho)\right]. \quad \blacksquare \end{aligned}$$

Lemma 2.4. *Let ξ_i be i.i.d. $\mathcal{N}(0,1)$ and $\mathcal{H}$ be a set of ordered multipliers. Then for any $\alpha \in (0, 1/4)$*

$$\begin{aligned} \mathbf{E}\max_{h\in\mathcal{H}}\left\{\pm\sum_{i=1}^n (h_i^2 - 2h_i)(\xi_i^2 - 1) - \alpha\sum_{i=1}^n h_i^2\right\} &\le \frac{K}{\alpha}, \\ \mathbf{E}\max_{h\in\mathcal{H}}\left\{\sum_{i=1}^n (1-h_i)^2\xi_i\mu_i - \alpha\sum_{i=1}^n (1-h_i)^2\mu_i^2\right\} &\le \frac{K}{\alpha}, \end{aligned}$$

where K is a generic constant.

Proof. It follows from Lemma 2 in [8].

Lemma 2.5. *Let*

$$\hat{h}^\epsilon = \max\left\{h : \left[\bar{r}(Y,\hat{\mu}^h) - \bar{r}^{\mathcal{H}}(Y)\right] \le 2\beta\epsilon\sigma^2\left[\|h\|^2 - \|\hat{h}\|^2\right] + 2\beta\sigma^2\right\}, \quad (2.23)$$

where $\epsilon \in (0, 1/(2\beta))$ and $\bar{r}^{\mathcal{H}}(Y) = \min_{h\in\mathcal{H}} \bar{r}(Y,\hat{\mu}^h)$. Then

$$\sqrt{\mathbf{E}_\mu\|\hat{h}^\epsilon\|^2} \le \sqrt{\frac{r^{\mathcal{H}}(\mu)}{(1-2\beta\epsilon)\sigma^2}} + \frac{\sqrt{1+2\beta}}{1-2\beta\epsilon}\sqrt{K}. \quad (2.24)$$

Proof. By the definition of $\bar{r}(Y, \hat{\mu}^h)$, see also (1.11) and (2.23), we get

$$\begin{aligned}\hat{h}^\epsilon =& \max\Big\{h:\ \|(1-h)\cdot\mu\|^2 + \sigma^2(1-2\beta\epsilon)\|h\|^2 \\ &+ 2\sigma\sum_{i=1}^n (1-h_i)^2\mu_i\xi_i + \sigma^2\sum_{i=1}^n (h_i^2-2h_i)(\xi_i^2-1) \\ \le& \|(1-\hat{h})\cdot\mu\|^2 + \sigma^2(1-2\beta\epsilon)\|\hat{h}\|^2 \\ &+ 2\sigma\sum_{i=1}^n (1-\hat{h}_i)^2\mu_i\xi_i + \sigma^2\sum_{i=1}^n (\hat{h}_i^2-2\hat{h}_i)(\xi_i^2-1) + 2\beta\sigma^2\Big\}.\end{aligned}$$

Let us fix some $\gamma, \gamma' > 0$. Then we can rewrite the above equation as follows:

$$\begin{aligned}\hat{h}^\epsilon = \max\Big\{h:\ & \sigma^2(1-2\beta\epsilon-\gamma)\|h\|^2 + 2\sigma\sum_{i=1}^n (1-h_i)^2\mu_i\xi_i + \|(1-h)\cdot\mu\|^2 \\ &+\sigma^2\sum_{i=1}^n (h_i^2-2h_i)(\xi_i^2-1) + \gamma\sigma^2\|h\|^2 \\ &\le (1+\gamma')\|(1-\hat{h})\cdot\mu\|^2 + \sigma^2(1-2\beta\epsilon+\gamma')\|\hat{h}\|^2 \\ &+2\sigma\sum_{i=1}^n (1-\hat{h}_i)^2\mu_i\xi_i - \gamma'\|(1-\hat{h})\cdot\mu\|^2 \\ &+\sigma^2\sum_{i=1}^n (\hat{h}_i^2-2\hat{h}_i)(\xi_i^2-1) - \gamma'\sigma^2\|\hat{h}\|^2 + 2\beta\sigma^2\Big\}.\end{aligned}$$

Therefore

$$\begin{aligned}\hat{h}^\epsilon \le \tilde{h}^\epsilon \stackrel{\text{def}}{=} \max\Big\{h:\ & \sigma^2(1-2\beta\epsilon-\gamma)\|h\|^2 \\ &+ \min_{g\in\mathcal{H}}\Big[2\sigma\sum_{i=1}^n (1-g_i)^2\mu_i\xi_i + \|(1-g)\cdot\mu\|^2\Big] \\ &+ \min_{g\in\mathcal{H}}\Big[\sigma^2\sum_{i=1}^n (g_i^2-2g_i)(\xi_i^2-1) + \gamma\sigma^2\|g\|^2\Big] \\ \le& (1+\gamma')\|(1-\hat{h})\mu\|^2 + (1-2\beta\epsilon+\gamma')\sigma^2\|\hat{h}\|^2 \\ &+ \max_{g\in\mathcal{H}}\Big[2\sigma\sum_{i=1}^n (1-g_i)^2\mu_i\xi_i - \gamma'\|(1-g)\cdot\mu\|^2\Big] \\ &+ \max_{g\in\mathcal{H}}\Big[\sigma^2\sum_{i=1}^n (g_i^2-2g_i)(\xi_i^2-1) - \gamma'\sigma^2\|g\|^2\Big] + 2\beta\sigma^2\Big\}.\end{aligned}$$

Next, bounding *max* and *min* in this equation with the help of Lemma 2.4, we arrive at

$$
\begin{aligned}
&(1-2\beta\epsilon)\sigma^2\mathbf{E}_\mu\|\tilde{h}^\epsilon\|^2 - \frac{K\sigma^2}{\gamma} - \gamma\sigma^2\mathbf{E}_\mu\|\tilde{h}^\epsilon\|^2 - K\sigma^2 \\
&\quad\le \mathbf{E}_\mu\|(1-\hat{h})\cdot\mu\|^2 + (1-2\beta\epsilon)\sigma^2\mathbf{E}_\mu\|\hat{h}\|^2 + \frac{K\sigma^2}{\gamma'} \\
&\qquad + \gamma'\mathbf{E}_\mu\|(1-\hat{h})\cdot\mu\|^2 + \frac{K\sigma^2}{\gamma'} + \gamma'\sigma^2\mathbf{E}_\mu\|\hat{h}\|^2 + 2\beta\sigma^2.
\end{aligned}
$$

Maximizing the left-hand side in $\gamma \in (0, 1/4)$ and minimizing the right-hand side in $\gamma' \in (0, 1/4)$, we obtain with a simple algebra

$$
\begin{aligned}
&\sigma^2\left[\sqrt{(1-2\beta\epsilon)\mathbf{E}_\mu\|\tilde{h}^\epsilon\|^2} - \left(\frac{K}{1-2\beta\epsilon}\right)^{1/2}\right]^2 \\
&\quad\le \left[\sqrt{\mathbf{E}_\mu\|(1-\hat{h})\cdot\mu\|^2 + \sigma^2\mathbf{E}_\mu\|\hat{h}\|^2} + \sigma\sqrt{K}\right]^2 + \frac{2\beta\epsilon K\sigma^2}{1-2\beta\epsilon} + (2\beta+K)\sigma^2.
\end{aligned}
$$

Combining this equation with $\|\hat{h}^\epsilon\|^2 \le \|\tilde{h}^\epsilon\|^2$ we arrive at

$$
\begin{aligned}
\sqrt{(1-2\beta\epsilon)\mathbf{E}_\mu\|\hat{h}^\epsilon\|^2} \le \sigma^{-1}&\sqrt{\mathbf{E}_\mu\|(1-\hat{h})\cdot\mu\|^2 + \sigma^2\mathbf{E}_\mu\|\hat{h}\|^2} \\
&+ \frac{\sqrt{1+2\beta\epsilon}}{\sqrt{1-2\beta\epsilon}}\sqrt{K} + \sqrt{2\beta+K}.
\end{aligned}
\tag{2.25}
$$

To control the expectation at the right-hand side in (2.25), we note that for any given $g \in \mathcal{H}$ the following inequality

$$
\sum_{i=1}^n [1-\hat{h}_i]^2 Y_i^2 + 2\sigma^2\sum_{i=1}^n \hat{h}_i \le \sum_{i=1}^n [1-g_i]^2 Y_i^2 + 2\sigma^2\sum_{i=1}^n g_i
$$

holds. Denoting for brevity

$$
\rho(h) \stackrel{\text{def}}{=} \|(1-h)\cdot\mu\|^2 + \sigma^2\|h\|^2,
$$

we rewrite this equation as follows:

$$
\begin{aligned}
&\rho(\hat{h}) + 2\sigma\sum_{i=1}^n (1-\hat{h}_i)^2\mu_i\xi_i + \sigma^2\sum_{i=1}^n(\hat{h}_i^2 - 2\hat{h}_i)(\xi_i^2-1) \\
&\quad\le \rho(g) + 2\sigma\sum_{i=1}^n (1-g_i)^2\mu_i\xi_i + \sigma^2\sum_{i=1}^n (g_i^2 - 2g_i)(\xi_i^2-1).
\end{aligned}
$$

So, for any $\gamma > 0$, we get with this equation and Lemma 2.4

$$\begin{aligned}\mathbf{E}_\mu \rho(\hat{h}) \leq & \rho(g) + \gamma \mathbf{E}_\mu \rho(\hat{h}) \\ & + 2\sigma \mathbf{E}_\mu \max_{h \in \mathcal{H}} \Big[- \sum_{i=1}^n (1-h_i)^2 \mu_i \xi_i - \frac{\gamma}{2\sigma} \sum_{i=1}^n (1-h_i)^2 \mu_i^2 \Big] \\ & + \sigma^2 \mathbf{E}_\mu \max_{h \in \mathcal{H}} \Big[\sum_{i=1}^n [2h_i - h_i^2](\xi_i^2 - 1) - \gamma \sum_{i=1}^n h_i^2 \Big] \\ \leq & \rho(g) + \frac{K\sigma^2}{\gamma} + \gamma \mathbf{E}_\mu \rho(\hat{h}).\end{aligned}$$

Next, minimizing the right-hand side in $g \in \mathcal{H}$ and $\gamma \in (0, 1/4)$, we obtain

$$\mathbf{E}_\mu \rho(\hat{h}) \leq r^{\mathcal{H}}(\mu) + 2\sigma \big[K \mathbf{E}_\mu \rho(\hat{h}) \big]^{1/2} + K\sigma^2$$

or, equivalently,

$$\Big\{ \big[\mathbf{E}_\mu \rho(\hat{h}) \big]^{1/2} - \sqrt{K}\sigma \Big\}^2 \leq r^{\mathcal{H}}(\mu) + K\sigma^2.$$

This yields obviously

$$\Big\{ \mathbf{E}_\mu \Big[\|(1-\hat{h}) \cdot \mu\|^2 + \sigma^2 \|\hat{h}\|^2 \Big] \Big\}^{1/2} \leq \sqrt{r^{\mathcal{H}}(\mu)} + 2\sigma\sqrt{K}.$$

Finally, substituting this inequality in (2.25), we get (2.24). ■

2.2 Proof of Theorem 1.3

By (2.11) we have by

$$\begin{aligned}\log[w_h(Y)] = & \frac{1}{2\sigma^2\beta} \bar{r}(Y, \hat{\mu}^{\hat{h}}) - \frac{1}{2\sigma^2\beta} \bar{r}(Y, \hat{\mu}^h) + \log \pi_h \\ & - \log \Big\{ \sum_{g \in \mathcal{H}} \pi_g \exp \Big[- \frac{\bar{r}(Y, \hat{\mu}^g) - \bar{r}(Y, \hat{\mu}^{\hat{h}})}{2\beta\sigma^2} \Big] \Big\},\end{aligned}$$

where $\hat{h} = \arg\min_{h \in \mathcal{H}} \bar{r}(Y, \hat{\mu}^h)$. Therefore

$$\begin{aligned}\sum_{h \in \mathcal{H}} w_h(Y) \bar{r}(Y, \hat{\mu}^h) = & \bar{r}(Y, \hat{\mu}^{\hat{h}}) + 2\beta\sigma^2 \sum_{h \in \mathcal{H}} w_h(Y) \log \frac{\pi_h}{w_h(Y)} \\ & - 2\beta\sigma^2 \log \Big\{ \sum_{g \in \mathcal{H}} \pi_g \exp \Big[- \frac{\bar{r}(Y, \hat{\mu}^g) - \bar{r}(Y, \hat{\mu}^{\hat{h}})}{2\beta\sigma^2} \Big] \Big\}.\end{aligned} \tag{2.26}$$

We begin to control the right-hand side at (2.26) with the last term. Ordering the elements in $\mathcal{H}$, we obtain by (2.2)

$$\begin{aligned}
&\log\Big\{\sum_{g\in\mathcal{H}}\pi_g\exp\Big[-\frac{\bar r(Y,\hat\mu^g)-\bar r(Y,\hat\mu^{\hat h})}{2\beta\sigma^2}\Big]\Big\}\\
&\quad\ge\log\Big\{\sum_{g\ge\hat h}\pi_g\exp\Big[-\frac{\bar r(Y,\hat\mu^g)-\bar r(Y,\hat\mu^{\hat h})}{2\beta\sigma^2}\Big]\Big\}\\
&\quad=\log\Big\{\sum_{g\ge\hat h}\pi_g\exp\Big[-\frac{\|(1-g)\cdot Y\|^2-\|(1-\hat h)\cdot Y\|^2}{2\beta\sigma^2}\\
&\qquad-\frac{1}{\beta}\sum_{i=1}^n[g_i-\hat h_i]\Big]\Big\}\ge\log\Big\{\sum_{g\ge\hat h}\pi_g\exp\Big[-\frac{1}{\beta}\sum_{i=1}^n[g_i-\hat h_i]\Big]\Big\}\ge 0.
\end{aligned}\tag{2.27}$$

Our next step is to bound from above the second term at the right-hand side of Equation (2.26). We note that for all $h>\hat h^\epsilon$, where $\hat h^\epsilon$ is defined by (2.23),

$$\big[\bar r(Y,\hat\mu^h)-\bar r^{\mathcal{H}}(Y)\big]\ge 2\beta\epsilon\sigma^2\big[\|h\|^2-\|\hat h\|^2\big]+2\beta\sigma^2.$$

Let

$$q_h=\exp\Big[-\frac{\bar r(Y,\hat\mu^h)-\bar r^{\mathcal{H}}(Y)}{2\beta\sigma^2}\Big]$$

and

$$\mathcal{G}=\big\{h\in\mathcal{H}:\|\hat h\|_1\le\|h\|_1<\|\hat h\|_1+1\big\}.$$

By Condition (1.16) we have that for all $h\ge\hat h^\epsilon$

$$q_h\le\exp\Big[-\frac{2\sigma^2\beta K_\circ\epsilon(\|h\|_1-\|\hat h\|_1)}{2\beta\sigma^2}-1\Big]=\exp\big[-K_\circ\epsilon(\|h\|_1-\|\hat h\|_1)-1\big].$$

So, we obtain with Lemma 2.3, Jensen's inequality, and (2.3)

$$\begin{aligned}
\mathbf{E}_\mu\sum_{h\in\mathcal{H}}w_h(Y)\log\frac{\pi_h}{w_h(Y)}&\le\mathbf{E}_\mu\log\Big\{\sum_{h\le\hat h^\epsilon}\pi_h+\exp[R(K_\circ\epsilon)]\Big\}\\
&\le\mathbf{E}_\mu\log\Big\{1+\frac{\|\hat h^\epsilon\|^2}{\beta K_\circ}+\exp[R(K_\circ\epsilon)]\Big\}\\
&\le\log\Big\{1+\frac{\mathbf{E}_\mu\|\hat h^\epsilon\|^2}{\beta K_\circ}+\mathbf{E}_\mu\exp[R(K_\circ\epsilon)]\Big\}\\
&\le 2\log\Big\{1+\Big[\frac{\mathbf{E}_\mu\|\hat h^\epsilon\|^2}{\beta K_\circ}\Big]^{1/2}+\big\{\mathbf{E}_\mu\exp[R(K_\circ\epsilon)]\big\}^{1/2}\Big\},
\end{aligned}\tag{2.28}$$

where $R(\cdot)$ is defined by (2.13).

Note that similar to (2.27) it can be checked easily that

$$q_h \geq \exp\left(-\frac{1}{\beta}\right),\ h \in \mathcal{G}$$

and thus it follows immediately from the definition of $\mathcal{G}$ and (2.4)-(2.5) that

$$\sum_{h\in\mathcal{G}} \pi_h q_h \geq \exp\left(-\frac{1}{\beta}\right) \sum_{h\in\mathcal{G}} \pi_h \geq \frac{1}{2\beta} \exp\left(-\frac{2}{\beta}\right)$$

and hence

$$R(K_\circ \epsilon) \leq \frac{C}{2\epsilon} + \frac{1}{2}\log\frac{C}{\epsilon}.$$

With this equation, bounding from above $\sqrt{\mathbf{E}_\mu \|\hat{h}^\epsilon\|^2}$ with the help of (2.24), we get from (2.26)–(2.28)

$$\begin{aligned}\mathbf{E}_\mu \sum_{h\in\mathcal{H}} w_h(Y)\bar{r}(Y,\hat{\mu}^h) \leq \mathbf{E}_\mu \bar{r}^{\mathcal{H}}(Y) + 4\beta\sigma^2 \log\Bigg\{ &\frac{1}{\sqrt{1-2\beta\epsilon}}\sqrt{\frac{r^{\mathcal{H}}(\mu)}{\sigma^2\beta K_\circ}} \\ &+ \frac{\sqrt{1+2\beta}}{1-2\beta\epsilon}\sqrt{K} + \exp\left[\frac{C}{2\epsilon} + \frac{1}{2}\log\frac{C}{\epsilon}\right]\Bigg\}.\end{aligned} \tag{2.29}$$

To finish the proof of the theorem, it remains to minimize the right hand side at this equation in ϵ. Assuming $\epsilon \leq 1/(3\beta)$, we obtain

$$\begin{aligned}&\frac{1}{\sqrt{1-2\beta\epsilon}}\sqrt{\frac{r^{\mathcal{H}}(\mu)}{\sigma^2\beta K_\circ}} + \frac{\sqrt{1+2\beta}}{1-2\beta\epsilon}\sqrt{K} + \exp\left[\frac{C}{2\epsilon} + \frac{1}{2}\log\frac{C}{\epsilon}\right] \\ &\quad\leq \sqrt{\frac{r^{\mathcal{H}}(\mu)}{\sigma^2\beta K_\circ}} + \frac{2\beta\epsilon}{\sqrt{1-2\beta\epsilon}}\sqrt{\frac{r^{\mathcal{H}}(\mu)}{\sigma^2\beta K_\circ}} + \frac{\sqrt{1+2\beta}}{1-2\beta\epsilon}\sqrt{K} \\ &\qquad + \exp\left[\frac{C}{2\epsilon} + \frac{1}{2}\log\frac{C}{\epsilon}\right] \\ &\quad\leq \sqrt{\frac{r^{\mathcal{H}}(\mu)}{\sigma^2\beta K_\circ}} + 3\sqrt{(1+2\beta)K} + 4\beta\epsilon\sqrt{\frac{r^{\mathcal{H}}(\mu)}{\sigma^2\beta K_\circ}} \\ &\qquad + \exp\left[\frac{C}{2\epsilon} + \frac{1}{2}\log\frac{C}{\epsilon}\right].\end{aligned} \tag{2.30}$$

Let

$$\Psi(x) = \min_{\epsilon\in[0,1/(3\beta)]}\left[\epsilon x + \sqrt{\frac{C}{\epsilon}}\exp\left(\frac{C}{2\epsilon}\right)\right].$$

It is clear that $\Psi(0)$ is bounded from above. It is also easy to check with $\epsilon = 4C/\log(x)$ that for any $x \geq 2$

$$\Psi(x) \leq \frac{4Cx}{\log(x)} + \frac{\sqrt{x\log(x)}}{2} \leq \frac{Cx}{\log(x)}.$$

So, combining (2.29) and (2.30) with Lemma 2.2, we complete the proof of (1.17) since $\mathbf{E}_\mu \bar{r}^{\mathcal{H}}(Y) \leq r^{\mathcal{H}}(\mu)$. ■

References

[1] Alquier, P. and Lounici, K. (2011). Pac-bayesian theorems for sparse regression estimation with exponential weights. *Electronic Journal of Statistics* **5**, 127–145.

[2] Arias-Castro, E. & Karim Lounici, K Variable Selection with Exponential Weights and l_0-Penalization. arXiv:1208.2635

[3] Akaike, H. (1973). Information theory and an extension of the maximum likelihood principle *Proc. 2nd Intern. Symp. Inf. Theory*. 267–281. MR0483125

[4] Catoni, O. (2004). *Statistical learning theory and stochastic optimization.* Lectures Notes in Math. **1851** Springer-Verlag, Berlin.

[5] Dalayan, A. and Salmon J. (2011). Sharp oracle inequalities for aggregation of affine estimators. *arXiv:1104.3969v2 [math.ST].*

[6] Demmler, A. and Reinsch, C. (1975). Oscillation matrices with spline smoothing. *Numerische Mathematik.* **24**, 375–382.

[7] Engl, H.W., Hanke, M., and Neubauer, A. (1996). *Regularization of Inverse Problems. Mathematics and its Applications, 375.* Kluwer Academic Publishers Group. Dordrecht.

[8] Golubev, Yu. (2010). On universal oracle inequalities related to high dimensional linear models. *Ann. Statist.* **38**, No. 5, 2751–2780.

[9] Golubev, G. (2012). Exponential weighting and oracle inequalities for projection estimates. *Problems of Information Transmission*, **48**, No. 3, 269–280.

[10] Green, P. J. and Silverman, B. W. (1994). *Nonparametric Regression and Generalized Linear Models. A roughness penalty approach*, Chapman and Hall.

[11] Juditsky, A. and Nemirovski, A. (2000). Functional aggregation for nonparametric regression. *Ann. Statist.* **28**, 681–712.

[12] Kneip, A. (1994). Ordered linear smoothers. *Annals of Stat.* **22**, 835–866.

[13] Lecué, G. (2007). Simultaneous adaptation to the margin and to complexity in classification. *Ann. Statist.* **35** 1698-1721. MR2351102

[14] Leung, G. and Barron, A. (2006). Information theory and mixing least-squares regressions. *IEEE Transactions on Information Theory* **52**, No. 8, 3396–3410.

[15] Mallows, C. L. (1973). Some comments on C_p *Technometrics* **15**, 661–675.

[16] Nemirovski, A. (2000). *Topics in non-parametric statistics.* Lectures Notes in Math. **1738** Springer-Verlag, Berlin.

[17] Nussbaum, M. (1985) Spline smoothing in regression models and asymptotic efficiency in L_2, *The Annals of Statist.* **13**, No. 3, 984-997.

[18] Rigollet, P. and Tsybakov, A. B. (2007). Linear and convex aggregation of density estimators. *Math. Methods Statist.* **16**, 260–280.

[19] Rigolet, Ph. and Tsybakov, A. (2012). Sparse estimation by exponential weighting. *Statistical Science*, **27**, No. 4, 558–575.

[20] Speckman, P. (1985). Spline smoothing and optimal rates of convergence in nonparametric regression. *Ann. Statist.* **13** 970–983. MR0803752

[21] Stein, C. (1981). Estimation of the mean of a multivariate normal distribution. *Ann. of Statist.*, **9**, 1135-1151.

[22] Tikhonov, A. N. and Arsenin, V. A. (1977). *Solution of Ill-posed Problems. Translated from the Russian. Preface by translation editor Fritz John. Scripta Series in Mathematics.* V. H. Winston & Sons, Washington, D.C.: John Wiley & Sons, New York.

[23] Wahba, G. (1990). *Spline Models for Observational Data.* SIAM, Philadelphia.

[24] Yang, Y. (2000). Combining different procedures for adaptive regression. *J. Multivariate Anal.* **74**, 135–161.